

Regression Analysis with Response-biased Sampling

Kani Chen

Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

Yuanyuan Lin

Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong

Yuan Yao

Department of Mathematics, Hong Kong Baptist University, Hong Kong

Chaoxu Zhou

Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

Abstract: Response-biased sampling, in which samples are drawn from a population according to the values of the response variable, is common in biomedical, epidemiological, economic and social studies. In particular, the complete observations in data with censoring, truncation or missing covariates can be regarded as response-biased sampling under certain conditions. This paper proposes to use transformation models, known as the generalized accelerated failure time model in econometrics, for regression analysis with response-biased sampling. With unknown error distribution, the transformation models are broad enough to cover linear regression models, the Cox's model and the proportional odds model as special cases. To the best of our knowledge, except for the case-control logistic regression, there is no report in the literature that a prospective estimation approach can work for biased sampling without any modification. We prove that the maximum rank correlation estimation is valid for response-biased sampling and establish its consistency and asymptotic normality. Unlike the inverse probability methods, the proposed method of estimation does not involve the sampling probabilities, which are often difficult to obtain in practice. Without the need of estimating the unknown transformation function or the error distribution, the proposed method is numerically easy to implement with the Nelder-Mead simplex algorithm, which does not require convexity or continuity. We propose an inference procedure using random weighting to avoid the complication of density estimation when using the plug-in rule for variance estimation. Numerical studies with supportive evidence are presented. Applications are illustrated with the Forbes Global 2000 data and the Stanford heart transplant data.

Key words and phrases: General transformation model; Maximum rank correlation; Random weighting; Response-biased sampling.

1. Introduction

Response-biased sampling is commonly used in biomedical, epidemiological, financial and social studies. In response-biased sampling, observations are taken according to the values of the responses. Specifically, let (Y^*, X^*) and (Y, X) represent the pair of response and covariates in the population and in the sample, respectively. In a response biased sampling, the conditional distribution of X given Y is the same as that of X^* given Y^* . Throughout the paper, we denote the observations as $(Y_i, X_i), i = 1, \dots, n$, which are independent and identically distributed. Data collected using response-biased sampling schemes are likely to contain more information relevant to one's interest than using prospective sampling. Such retrospective sampling is useful in clinical studies for its effectiveness and its saving duration and costs. For example, in a study of possible dependence of levels of hypertension (response) on those of sodium intake (covariate), sampling from patients in a hospital, which can be regarded as response-biased sampling, would be more effective than from general public as the latter has much smaller proportion of people with hypertension. Moreover, a typical example of sampling with selection bias in economic and social studies is that the wage is only observed for the employed people. The statistical analysis of biased sampling has received considerable attention in the past decades. Case-control or choice-based sampling, which is a special case of response-biased sampling, has been extensively studied in the literature; see Anderson (1972), Manski and Lerman (1977), Prentice and Pyke (1979), Breslow and Day (1980), Cosslet (1981), Scott and Wild (1986, 1997), Manski (1993), etc. There are other studies on biased sampling data, involving semiparametric and parametric models; see Hausman and Wise (1981), Jewell (1985), Bickel and Ritov (1991), Wang (1996), Lawless et al. (1999), Chen (2001), Tsai (2009), Luo and Tsai (2009), Luo et al. (2009), among others. In statistical analysis of biased sampling, one of the celebrated findings is that the prospective estimating equation is still valid for case-control logistic regression; see Anderson(1977) and Prentice and Pyke (1979). However, in general, estimating equations based on prospective sampling will be invalid for biased sampling and modifications using, for example, inverse probability

method is necessary. This paper shows, for general transformation model, a rank estimation method based on prospective sampling still applies, without any modification, to response-biased sampling.

Regression analysis with response-biased sampling is generally associated with the fitted model. In particular, the estimation of the parameter of interest with biased sampling usually relies on the model assumptions, such as the inverse probability method and the pseudo-likelihood method; see Binder (1992), Lin (2000), Wang (1996) and Tsai (2009). Recently, nonparametric tests and estimation for right censored data with biased sampling can be found in Ning, Qin and Shen (2010) and Huang and Qin (2011). Moreover, a novel approach to analyze length-biased data with semiparametric transformation and accelerated failure time models has been developed by Shen, Ning and Qin (2009). In this paper, we consider a class of transformation models with response-biased sampling, under which an unknown monotonic transformation of the response is linearly related to the covariates with an unspecified error distribution. The transformation models are also called the generalized accelerated failure time (GAFT) model in econometrics. This class of regression models includes many popular models, such as the proportional hazards model, the proportional odds model as well as accelerated failure time models or linear models. Furthermore, the response-biased sampling that we consider can be viewed as a special case of the celebrated Heckman model; see Heckman (1977, 1979). The Heckman model assumes an outcome linear regression model and a probit selection model. We consider more general transformation models and assume the "selectivity/observability" solely depends on the value of the response variable. In the case analysis of wage, we assume the chance that a potential job is taken only depends on the wage offered. The proposed estimating method does not depend on the specification of the sampling probabilities, unlike the well known Heckman correction. We note that there is a rich literature on linear transformation models with a known error distribution; see, for example, Dabrowska and Doksum (1988), Cheng et al. (1995, 1997), Chen et al. (2002) and Zeng and Lin (2007). However, their reported methods cannot be directly applied to transformation models with an unknown error distribution. Similarly, the case-control logistic regression method in Anderson (1977) and Prentice and Pyke (1979) which works only for a special model,

cannot be generalized directly and modification using, for example, the inverse probability method is inevitable. However, for the inverse probability method, its validity requires correct prospective mean zero estimating equations, correct specifications of sampling probabilities, and the sampling probability must be positive for every value in the range of the response.

In view of the importance of the response-biased sampling designs as well as transformation models, an easy-to-implement estimation methodology, with an advantage over the existing methods in terms of generality, is worth pursuing. Note that the conventional methods, such as the least squares (LS) or the least absolute deviations (LAD) cannot be directly applied to response-biased sampling, because the zero mean and the zero median assumptions do not hold anymore. The maximum rank correlation (MRC) estimate, originated from Han (1987) for prospective studies, is based on the rank correlation (Kendall's τ) between two variables. For illustration, consider a simple linear regression model

$$Y_i = \beta' X_i + \epsilon_i, \quad 1 \leq i \leq n,$$

where (Y_i, X_i, ϵ_i) are independent and identically distributed (*i.i.d.*) copies of (Y, X, ϵ) . The idea of the MRC estimation is to maximize the rank correlation between Y_i and $\beta' X_i$ with respect to β . Heuristically, given that $\beta' X_i > \beta' X_j$, it is more likely that $Y_i > Y_j$ than otherwise. In other words, the rank of Y_i and the rank of $\beta' X_i$ are positively correlated. A number of studies on MRC have been conducted. Sherman (1993) proved its \sqrt{n} -consistency and asymptotic normality and Khan and Tamer (2004) extended this method to semiparametric models with censoring by proposing the partial rank (PR) estimator. A smoothed partial rank (SPR) estimator is then considered in Song et al. (2007) for transformation models with censoring.

Inspired by the fact that response-biased sampling would not change the positive correlation between the ranks of the responses and explanatory variables, this article offers an estimator based on MRC for transformation models with response-biased sampling. The proposed estimation does not rely on any further model assumption. It works equally well regardless of what the monotonic transformation is, as the MRC estimate only depends on the ranks of responses. The estimation of the transformation function, which is likely to be quite complex and computationally burdensome, is not required. The proposed method is easy

to implement and computationally straightforward with the help of Nelder-Mead simplex direct search. It is quite well known that the Nelder-Mead simplex algorithm does not require convexity or continuity; see Nelder and Mead (1965). Note that the MRC objective function is a U-statistic. In order to avoid estimating the covariance matrix, we propose to use a random weighting resampling scheme for inference. In addition, since prospective sampling can be regarded as a special case of response-biased sampling, the proposed estimation is valid for prospective sampling.

We describe the model in section 2. The proposed estimation and its inference with theoretical justification are presented in section 3. A simulation study with supportive evidence is given in section 4. In section 5, our method is applied to the Forbes Global 2000 data set and the Stanford heart transplant data set. The paper concludes with a remark in section 6. All proofs are deferred to the Appendix.

2. Model description

Let (Y^*, W^*) be a pair of a response and a $(d + 1)$ -dimensional vector of covariates in the population. Assume that the response depends on the covariates according to the transformation model

$$H(Y^*) = \theta_0' W^* + \epsilon^*, \quad (2.1)$$

where $H(\cdot)$ is an unknown monotonically increasing function, ϵ^* is the error, independent of W^* , with unspecified distribution, and θ_0 is a $(d + 1)$ -dimensional vector of regression coefficients. When $H(\cdot)$ is the identity function, model (2.1) becomes a linear regression model. When ϵ^* follows the extreme-value distribution and the standard logistic distribution, the resulting model is the proportional hazards model and the proportional odds model, respectively. In prospective studies, model (2.1) has been extensively studied in the literature. Note that θ_0 in model (2.1) is not identifiable, meaning that θ_0 is not uniquely defined. To avoid unidentifiability, one may restrict $\|\theta_0\| = 1$. Without loss of generality, we choose to fix the first component of θ_0 to be 1. Then, $\theta_0 = (1, \beta_0')'$, where β_0 denotes the rest components. Accordingly, W^* can be decomposed into $W^* = (Z^*, X^*)$, where Z^* is the covariate corresponding to the fixed regression coefficient and X^* is the other d -dimensional covariate. Hence, model (2.1) can

be rewritten as

$$H(Y^*) = Z^* + \beta'_0 X^* + \epsilon^*. \quad (2.2)$$

Let (Y, Z, X) be the response and covariates following the distribution of response-biased sampling. The nature of response-biased sampling implies that, for any y , the conditional distribution of (X, Z) given $Y = y$ is the same as that of (X^*, Z^*) given $Y^* = y$. An alternative but equivalent definition of response biased sampling is by using a sampling index Δ . The pair of response and covariates, (Y^*, X^*) , are observed if and only if $\Delta = 1$. Then the response biased sampling is defined by the conditional independence of Δ and X^* given Y^* . And we can denote the observations as (Y_i^*, X_i^*, Δ_i) where $\Delta_i = 1$ for $i = 1, \dots, n$. With biased-sampling, Wang (1996) provided a novel pseudo-likelihood method for Cox's proportional hazards model. Recently, a pseudo-partial likelihood approach can be found in Tsai (2009). The existing methods for Cox's model with biased-sampling are conceptually appealing and have clear interpretation. To the best of our knowledge, no specific construction of regression analysis based on transformation models with response-biased sampling is available in the literature. In the next section, we propose a general estimation and inference procedure based on MRC for model (2.2) with response-biased sampling.

3. Estimation and inference

With response-biased sampling, the observations are (Y_i, Z_i, X_i) , $1 \leq i \leq n$, which are *i.i.d.* copies of (Y, Z, X) . Throughout the paper, $I(\cdot)$ is the indicator function. Similar to Han (1987), the rank correlation for response-biased sampling is defined as

$$U_n(\beta) = \sum_{i \neq j} I(Z_i + \beta' X_i > Z_j + \beta' X_j) I(Y_i > Y_j). \quad (3.1)$$

The MRC estimate is to maximize the rank correlation $U_n(\beta)$. Denote $\hat{\beta}_n$ as the maximizer of $U_n(\beta)$. Han (1987) and Sherman (1993) established the consistency and asymptotic normality of $\hat{\beta}_n$ with data from prospective sampling. However, with response sampling, it is not clear whether the large sample properties still hold.

The following theorem presents the consistency and asymptotic normality for $\hat{\beta}_n$ with response-biased sampling.

Theorem 1. *Under regularity conditions C1-C4 given in the Appendix, as $n \rightarrow \infty$,*

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow N(0, A^{-1}B(A^{-1})')$$

in distribution, where the explicit forms of A and B are given in the Appendix.

It is shown in the Appendix that the limiting covariance matrix of $\hat{\beta}_n$ involves the derivative of conditional expectation of the objective function, which could be quite difficult to estimate. To circumvent the difficulty, we propose a distributional approximation based on random weighting method by externally generating *i.i.d.* random variables. Let e_1, \dots, e_n be a sequence of *i.i.d.* nonnegative random variables with mean 1 and variance 1. Define

$$\tilde{U}_n(\beta) = \sum_{i \neq j} e_i e_j I(Z_i + \beta' X_i > Z_j + \beta' X_j) I(Y_i > Y_j) \quad (3.2)$$

and $\tilde{\beta}_n = \arg \max_{\beta \in \mathcal{B}} \tilde{U}_n(\beta)$. The distribution of $\sqrt{n}(\hat{\beta}_n - \beta_0)$ can be approximated by the resampling distribution of $\sqrt{n}(\tilde{\beta}_n - \hat{\beta}_n)$ when fixing the data $(Y_i, Z_i, X_i), 1 \leq i \leq n$.

Proposition. *Given $\{(Y_i, Z_i, X_i), 1 \leq i \leq n\}$, under regularity conditions C1-C4 in the Appendix, as $n \rightarrow \infty$,*

$$\sqrt{n}(\tilde{\beta}_n - \hat{\beta}_n) \rightarrow N(0, A^{-1}B(A^{-1})')$$

in distribution, which is the asymptotic distribution of $\sqrt{n}(\hat{\beta}_n - \beta_0)$.

The resampling method based on random weighting for the U-statistic objective function is well established in Jin (2001). We omit the proofs of the proposition here.

Remark 1. For the computation, the numerical minimization is straightforward with the Nelder-Mead simplex algorithm which does not require convexity or continuity. In the simulation, we use Nelder-Mead algorithm directly to search over a wide range of starting values in case there may exist local maximizers. Matlab code is available upon request. In addition, another slight problem is that, with large sample size or large dimension of covariates, the computation tends to be slower in simulation due to many replications. However, an algorithm proposed by Abrevaya (1999) which improves the complexity of computation for MRC from $O(n^2)$ to $O(n \log n)$ is available for large sample size. And a smoothed

approximation of the indicator function considered by Song et al. (2007) can be applied for large dimension of covariates. Overall, the proposed method has little difficulty in numerical implementation.

Remark 2. Note that our objective function $U_n(\beta)$ only depends on the responses through their orders which are not changed by the unknown monotonically increasing transformation $H(\cdot)$. Thus our estimate of β_0 is invariant of the transformation and estimating the unknown transformation $H(\cdot)$ can be avoided.

Remark 3. Response-biased sampling is related to truncated and censored data. With the presence of left-truncation, let X^* be covariates, Y^* be response and C^* be the left-truncation variable. Then, (X^*, Y^*, C^*) is observed if and only if $Y^* \geq C^*$, and the observation, denoted as (X, Y, C) accordingly, follows the conditional distribution of (X^*, Y^*, C^*) given $Y^* \geq C^*$. The observed pair of covariates and response, (X, Y) , can be treated as a special case of response-biased sampling, if C^* is independent of X^* and Y^* . Specifically, the conditional density of X given Y can be formally written as

$$f_{X|Y}(x|y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} = \frac{\int f_{(X,Y,C)}(x,y,c)dc}{\int f_{(Y,C)}(y,c)dc} = \frac{\int f_{(Y,C)}(y,c)f_{X|(Y,C)}(x|(y,c))dc}{\int f_{(Y,C)}(y,c)dc}.$$

The independence of C^* and (X^*, Y^*) gives

$$f_{X|(Y,C)}(x|(y,c)) = \frac{f_{(X^*,Y^*)}(x,y)f_{C^*}(c)/P(Y^* \geq C^*)}{f_{Y^*}(y)f_{C^*}(c)/P(Y^* \geq C^*)} = f_{X^*|Y^*}(x|y),$$

which is irrelevant with c . Thus, the conditional distribution of $X|Y$ is the same as that of $X^*|Y^*$. Similarly, for right-censored data with the censoring variable \tilde{C} independent of X^* and Y^* , denote the observation as (X, Y, δ) , where $X = X^*$, $Y = \min(Y^*, \tilde{C})$ and $\delta = I(Y^* \leq \tilde{C})$. Then, the conditional density is

$$f_{X|(Y,\delta=1)}(x|(y,\delta=1)) = \frac{f_{(X^*,Y^*)}(x,y)P(\tilde{C} \geq y)}{f_{Y^*}(y)P(\tilde{C} \geq y)} = f_{X^*|Y^*}(x|y).$$

Therefore the uncensored observations can be regarded as drawn from a response-biased sampling. Note that the partial rank method, which works for censoring data set, cannot be applied to truncation data. The our method works better in this view as it can handle a broad class of data types including left-truncation and right-censoring.

Remark 4. For data with missing covariates, the complete observations can be regarded as drawn from a response biased sampling, if the missing mechanism

is missing-at-random. This is because, by the definition of missing-at-random, the conditional distribution of the covariates given the response for the complete cases is the same as that for the observations with missing covariates and, as a result, also same as that in the population.

Remark 5. Conditions C3 in the Appendix is imposed to facilitate the proof of consistency. We assume that the error distribution has a twice differentiable density function with log-concavity. Although it looks somewhat restrictive, it includes a number of widely used distributions, for example, $N(0, \sigma^2)$ and Pareto family. Thus linear models with normal errors, Cox's model and the proportional odds model are included. With increasing technicalities, this condition might be loosened or dropped, as evidenced in our simulation results in section 4.

4. Simulation studies

Extensive simulation studies are conducted to examine the finite sample performance of the proposed method, which are presented in four parts. In the first part, we consider the linear model

$$Y = Z + X_1\beta_1 + X_2\beta_2 + \epsilon, \quad (4.1)$$

where $(\beta_1, \beta_2) = (1, -1)$, $Z \sim N(0, 1)$, and X_1 and X_2 follow a bivariate normal distribution with mean $(1, -0.5)$ and variance

$$\begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}.$$

Response-biased sampling are conducted with the five different schemes. In schemes 1, 2 and 5, the samples are restricted to $Y < -2$ or $Y > 4$, $Y > 2.5$ and $3.8 < Y < 4.2$, respectively. In scheme 3, the sampling probability is fixed as $\Phi(y - 2)$ for response value y . Scheme 4 is simply the prospective sampling. Four distributions for the error are used: double exponential distribution with parameter 1, the standard normal distribution, the standard extreme value distribution and the standard logistic distribution. The sample size is 200 and simulation results are based on 100 replications. The external random weights are generated from standard exponential distribution with 500 replications. For comparison, we also conduct simulation studies using inverse probability method with the same settings of the above five sampling schemes, in which the inverse sampling probabilities are the weights of the least square estimating equations; see Horvitz

and Thompson (1952). In Table 1, we present the bias of the estimates of the regression parameters β_1 and β_2 (BIAS), the empirical standard error (SE), the average of the estimated standard errors (SEE) and the 95% coverage probabilities (CP) with the proposed method. We also present the estimation results with the inverse probability method in Table 1.

INSERT TABLE 1 HERE

It can be seen from Table 1 that the proposed method works well with all different sampling schemes and error distributions. The estimated standard errors based on random weighting are close to the empirical standard errors in general. The proposed method offers more accurate and stable estimates compared with the inverse probability method for most of the examples. Except for sampling scheme 4 (prospective sampling) with normal and double exponential error distributions, the inverse probability method gives inaccurate estimates in general. This is mainly because the validity of the inverse probability method requires correctly specifying prospective mean zero estimating functions and positive sampling probabilities, that are easily violated. Overall, the first part of the simulation contains strong evidence of the superiority of the proposed method over the inverse probability method, in terms of both generality and flexibility.

The second part of the simulation is intended to show condition C3 may just be technical. Consider the model

$$Y = Z + \beta'X + \epsilon,$$

where $\beta = 1$, $Z \sim N(0, 1)$, $X \sim N(0, 1)$, ϵ follows the mixture of the standard normal distribution and a Bernoulli distribution with probability of success 0.5 and the mixture probabilities are (0.5, 0.5). The error distribution is not log-concave and thus does not satisfy condition C3. Samples with values of the response less than -1.5 or greater than 2.5 are drawn. The bias of the estimate is 0.0302. The empirical and estimated standard deviations are 0.1591 and 0.1479, respectively. The proposed method may still work without assuming the log-concavity of the error distribution.

The third part uses a rather extreme example to demonstrate that a biased sampling could be much more efficient than prospective sampling. Consider the

model

$$Y = Z + \beta'X + \epsilon,$$

where $\beta = 1$, $\epsilon \sim N(0, 10^{-4})$, $Z \sim N(0, 1)$, and X follows a distribution with density function

$$f_X(x) = \begin{cases} 5 * 10^{-5}, & \text{if } -105 \leq x \leq -5; \\ 9.99, & \text{if } -0.05 \leq x \leq 0.05; \\ 5 * 10^{-5}, & \text{if } 5 \leq x \leq 105. \end{cases}$$

For a response-biased sampling which only takes observations with responses larger than 4.5 or smaller than -4.5 , the mean and standard deviation of the estimates of β are 1.0004 and 0.0038, respectively. On the other hand, for the prospective sampling, the mean and standard deviation of the estimates of β are 1.0108 and 0.0576, respectively. The relative efficiency for the response-biased sampling versus the prospective sampling is 230. It indicates the possibility that response-biased sampling can be designed more efficiently than prospective sampling in terms of parameter estimation.

Overall, the results of simulation studies agree with the theory. Moreover, the consistency and asymptotic normality established in Theorem 1 might hold in more general scenarios, without the technical conditions.

5. Applications

In this section, we apply the proposed method to the Forbes Global 2000 data published in 2012 and the Stanford heart transplant data.

The first data set contains the profits, assets and market value for companies of the Forbes Global 2000. It is commonly known that profits is a measure of the financial performance of the companies and assets indicate the size of the companies. The purpose of this study is to analyze the relationship among market value, profits and assets of companies with the existing Forbes Global 2000 data. But the companies on the Forbes Global 2000 list, that are the biggest and most powerful companies in the world, is in fact a biased sampling data from the population. The sample size here $n = 2000$. We fit the transformation model to the data with covariates $X_1 = \text{assets}/250$, $Z = \text{profits}$ and response $Y = \text{market value}$ with the proposed method. For identifiability, we set the coefficient of Z to 1. The random weights are generated from the standard exponential distribution

with resampling times $N = 500$. The estimate of the coefficient of X_1 is 0.2912 and the estimated standard error is 0.0503.

Our second example pertains to the Stanford heart transplant data. Crowley and Hu (1977) reported information of 103 potential heart transplant recipients in the Stanford heart transplantation program consisting age, waiting time to transplantation, survival or censoring time from acceptance to the programme, and three mismatch scores from October 1967 to April 1974. During that time, 69 of the patients underwent the operation. Miller and Halpern (1982) reported the survival times, censored or uncensored in February 1980 of 184 patients who had received heart transplants. Similar to Miller and Halpern (1982), we consider the 152 patients whose $T5$ mismatch scores were not missing and survival times were not less than 10 days. Thus the observations in the study are left-truncated and right-censored. In view of remark 3 in section 3, the 97 complete observations can be treated as drawn from a response-biased sampling. We regress the *survival time* against *age* and age^2 with the transformation model. The coefficient of *age* is fixed to 1 and the random weights are generated from standard exponential distribution with 500 replications. Our proposed method gives the estimate of the coefficient of age^2 being -0.0152 with standard error 0.0031 and the 95% confidence interval $[-0.0213, -0.0091]$. To compare with the Cox's estimator presented in Miller and Halpern (1982), we consider the ratio of the coefficient of age^2 to the coefficient of *age*. The resulting estimate is -0.0161 with standard error 0.0016 and the 95% confidence interval $[-0.0191, -0.0130]$. It can be seen from both methods that the confidence interval, which does not cover zero, confirms the negative quadratic effect of the *age*. Note that the confidence interval of Cox's estimator is contained inside that obtained from the proposed method. This is mainly because the sample size from response-biased sampling is smaller and the transformation model is more general than the Cox's model. In addition, a comparison among different methods applying to an earlier published Standard heart transplant data set can be found in Khan and Tamer (2007), which also gives a similar estimating result to the MRC method.

6. Concluding remarks

This paper gives a general method of regression analysis based on the method of MRC for transformation models with response-biased sampling. Consistency

and asymptotic normality of the proposed estimator are proved theoretically. Simulation studies show that response-biased sampling gives a more efficient estimation than prospective sampling in certain situations, and the proposed estimator works well for a variety of sampling schemes and models. In addition, the nature of the MRC method implies that the estimation does not vary with different monotonic transformations, avoiding the estimation of the transformation functions. Furthermore, this method can be applied to more general models of the form

$$Y^* = D \cdot F(\theta_0' W^*, \epsilon^*), \quad (6.1)$$

where Y^* , W^* , θ_0 and ϵ^* are defined as in section 2. $D: \mathbb{R} \rightarrow \mathbb{R}$ is a non-degenerate, monotonic function and $F: \mathbb{R}^2 \rightarrow \mathbb{R}$ is strictly monotonic in each of the variables. Though we cannot separate the covariate term and the error term in this model, our estimation and inference procedure can still be applied as long as the monotonicity assumptions of the composite transformation $D \cdot F$ are valid.

References

- ABREVAYA, J. (1999). Computation of the maximum rank correlation estimator. *Economics letters* **62**, 279–285.
- ANDERSON, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.
- BICKEL, P. J. & RITOV, Y. (1991). Large sample theory of estimation in biased sampling regression models. *Ann. Statist.* **19**, 797–816.
- BINDER, D. A. (1992). Fitting Cox’s proportional hazards models from survey data. *Biometrika* **79**, 139–147.
- BRESLOW, N. E. & DAY, N. E. (1980). *The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.
- CHEN, K. (2001). Parametric models for response-biased sampling. *J. R. Statist. Soc. B.* **63**, 775–789.
- CHEN, K., JIN, Z. & YING, Z. (2002). Semiparametric analysis of transformation model with censored data. *Biometrika* **89**, 659–668.

- CHENG, S. C., WEI, L. J. & YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–845.
- CHENG, S. C., WEI, L. J. & YING, Z. (1997). Prediction of survival probabilities with semi-parametric transformation models. *J. Am. Statist. Assoc.* **92**, 227–235.
- COSSLET, S.R. (1981). Maximum likelihood estimate for choice-based samples. *Econometrica* **49**, 1289–1316.
- DABROWSKA, D. M. & DOKSUM, K. A. (1988). Estimation and testing in the two-sample generalized odds-rate model. *J. Am. Statist. Assoc.* **83**, 744–749.
- HAN, A. K. (1987). Non-parametric analysis of a generalized regression model. *J. Econometrics* **35**, 303–316.
- HAUSMAN, J. A. & WISE, D. A. (1981). Stratification on endogenous variables and estimation: the Gary Income Maintenance Experiment. In *Structural Analysis of Discrete Data: with Econometric Applications* (eds C. Manski and D. McFadden), pp. 364–391. Cambridge: Massachusetts Institute of Technology Press.
- HECKMAN, J. J. (1977). Sample selection bias as a specification error with an application to the estimation of labor supply functions. NBER working paper #172.
- HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–161.
- HORVITZ, D. G. & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* **47**, 663–685.
- HUANG, C. Y. & QIN, J. (2010). Nonparametric estimation for length-biased and right-censored data. *Biometrika* **98**, 177–186.
- JEWELL, N. (1985). Regression from stratified samples of dependent variables. *Biometrika* **72**, 11–21.

- JIN, Z., YING, Z. & WEI, L. J. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381–390.
- KHAN, S. & TAMER, E. (2007). Partial rank estimation of duration models with general forms of censoring. *J. Econometrics* **136**, 251–280.
- LAWLESS, J. F., KALBFLEISCH, J. D. & WILD, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. R. Statist. Soc. B.* **61**, 413–438.
- LIN, D. Y. (2000). On fitting Cox’s proportional hazards models to survey data. *Biometrika* **87**, 37–47.
- LUO, X., & TSAI, W. Y. (2009). Nonparametric estimation for right-censored length-biased data: a pseudo-partial likelihood approach. *Biometrika* **96**, 873–886.
- LUO, X., TSAI, W. Y. & XU, Q. (2009). Pseudo-partial likelihood estimators for the Cox regression model with missing covariates. *Biometrika* **96**, 617–633.
- MANSKI, C. F. (1993). The selection problem in econometrics and statistics. *Econometrika* (eds G. S. Maddala, C. R. Rao and H. D. Vinod), pp. 73–84. Amsterdam: North-Holland.
- MANSKI, C. F. & LERMAN, S. (1977). The estimation of choice probabilities from choice-based samples. *Econometrica* **45**, 1977–1988.
- MILLER, R. & HALPERN, J. (1982). Regression with censored data. *Biometrika* **69**, 521–531.
- NELDER, J. A. & MEAD, R. (1965). A simplex algorithm for function minimization. *Computer Journal* **7**, 308–313.
- NING, J., QIN, J. & SHEN, Y. (2010). Nonparametric tests for right-censored data with biased sampling. *J. R. Statist. Soc. B.* **72**, 609–630.
- PRENTICE, R. L. & PYKE, R. (1979). Logistic disease incidence models with case-control studies. *Biometrika* **66**, 403–411.

- SCOTT, A. J. & WILD, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *J. R. Statist. Soc. B.* **48**, 170–182.
- SHERMAN, R. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61**, 123–137.
- SHERMAN, R. (1994). Maximal Inequalities for Degenerate U-processes with Applications to Optimization Estimators. *Ann. Statist.* **22**, 439–459.
- SHEN, Y., NING, J. & QIN, J. (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *J. Am. Statist. Assoc.* **104**, 1192–1202.
- SONG, X., MA, S., HUANG, J. & ZHOU, X.H. (2007). A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics* **8**, 197–211.
- STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348–1360.
- STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053.
- TSAI, W. Y. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika* **96**, 601–615.
- VAN DER VAART, A. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag.
- WANG, M. C. (1996). Hazards regression analysis for length-biased data. *Biometrika* **83**, 343–354.
- ZENG, D. & LIN, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Statist. Soc. B.* **69**, 507–564.

Kani Chen, Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

E-mail: makchen@ust.hk

Yuanyuan Lin, Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong

E-mail: yy.lin@polyu.edu.hk

Yuan Yao, Department of Mathematics, Hong Kong Baptist University, Hong Kong

E-mail: yaoyuan@hkbu.edu.hk

Chaoxu Zhou, Department of Mathematics, Hong Kong University of Science and Technology, Hong Kong

E-mail: chaoxu.zhou@gmail.com

Appendix: Proof of Theorem 1

Consider the transformation model

$$H(Y^*) = \theta_0' W^* + \epsilon^*, \quad (7.1)$$

where $H(\cdot)$ is an unknown monotonically increasing function, ϵ^* is the error, independent of W^* , with unspecified distribution, and θ_0 is a $(d+1)$ -dimensional vector of regression coefficients. Accordingly, W^* can be decomposed into $W = (Z^*, X^*)$, where Z^* is the covariate corresponding to the fixed regression coefficient and X^* is the other d -dimensional covariate. Hence, the model can be rewritten as

$$H(Y^*) = Z^* + \beta_0' X^* + \epsilon^*.$$

We point out that the parameter estimation does not vary with different decompositions of covariates. Suppose that $(\tilde{Z}^*, \tilde{X}^*)$ is another composition of covariates. Then there exists a unique matrix P of full rank such that $(\tilde{Z}^*, \tilde{X}^{*'})' = P(Z^*, X^{*'})'$.

Suppose that $\tilde{\beta}'(\tilde{Z}^*, \tilde{X}^{*'})' = \beta'(Z^*, X^{*'})'$. Then by the uniqueness of linear representation, the relevant parameter must satisfy that $P\tilde{\beta} = \beta$. So if one parameter is uniquely determined in a d -dimensional linear space, the other parameter is also uniquely determined in a transformed d -dimensional linear space.

For easier explanation in the technical proof, we rewrite the transformation model (7.1) into

$$H(Y^*) = Z^* + \beta_0' X^* + \epsilon^*,$$

where we suppose the covariance decomposition satisfies that $\tilde{Z}^* := Z^* + \beta_0' X^*$ is irrelevant of X^* . Such a decomposition always exists since $\theta_0' W^*$ is a one-dimensional vector in a $(d+1)$ -dimensional linear space, so it has a d -dimensional orthogonal complement which can be defined as X^* . Furthermore, \tilde{Z}^* and X^* are supposed to be independent.

Suppose the regularity conditions hold:

- C1) The unknown parameter β lies in a bounded space $\mathbf{B} \subset \mathcal{R}^d$;
- C2) Both of Z^* and X^* have continuously differentiable density functions to the second order;
- C3) f_{ϵ^*} is log-concave (i.e., $\log f_{\epsilon^*}$ is concave);

C4) (Identifiability condition) $\xi(\beta) := (\beta - \beta_0)'(X_2^* - X_1^*) = 0$ almost surely if and only if $\beta = \beta_0$.

Consistency:

Define

$$g(\beta) = E[I\{Y_1 < Y_2\}I\{\beta X_1 + Z_1 < \beta X_2 + Z_2\}]$$

and

$$g_n(\beta) = \frac{1}{n^2 - n} \sum_{i \neq j} I\{Y_i < Y_j\} I\{\beta X_i + Z_i < \beta X_j + Z_j\}.$$

Step 1. We show that $g(\beta)$ has a unique maximum at $\beta = \beta_0$.

Write, for any $t_1 < t_2$,

$$\begin{aligned} & E[I\{Y_1 < Y_2\}I\{\beta X_1 + Z_1 < \beta X_2 + Z_2\} | Y_1 = t_1, Y_2 = t_2] \\ &= P(\beta X_1 + Z_1 < \beta X_2 + Z_2 | Y_1 = t_1, Y_2 = t_2) \\ &= P(\beta X_1^* + Z_1^* < \beta X_2^* + Z_2^* | Y_1^* = t_1, Y_2^* = t_2) \\ &= P(Z_1^* - Z_2^* < \beta X_2^* - \beta X_1^* | \beta_0 X_1^* + Z_1^* + \epsilon_1^* = H(t_1), \beta_0 X_2^* + Z_2^* + \epsilon_2^* = H(t_2)) \\ &= P(\tilde{Z}_1^* - \tilde{Z}_2^* < (\beta - \beta_0)(X_2^* - X_1^*) | \tilde{Z}_1^* + \epsilon_1^* = \tilde{t}_1, \tilde{Z}_2^* + \epsilon_2^* = \tilde{t}_2) \\ &= \frac{\int P(\xi(\beta) > s_1 - s_2) f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2}{\int f_{\tilde{Z}^*}(s) f_{\epsilon^*}(\tilde{t}_1 - s) ds \int f_{\tilde{Z}^*}(s) f_{\epsilon^*}(\tilde{t}_2 - s) ds}, \end{aligned} \quad (7.2)$$

where $\tilde{t}_i = H(t_i)$, $i = 1, 2$.

The denominator is irrelevant with β . The numerator will be proved to have a unique maximum at $\beta = \beta_0$. The numerator can be written as

$$\begin{aligned} & \frac{1}{2} \int [1 - \text{sgn}(s_1 - s_2) P(|\xi(\beta)| < |s_1 - s_2|)] \\ & \quad f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2 \\ &= \frac{1}{2} \int f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2 + \Pi(\beta) \end{aligned}$$

where

$$\begin{aligned} \Pi(\beta) &= -\frac{1}{2} \int \text{sgn}(s_1 - s_2) P(|\xi(\beta)| < |s_1 - s_2|) \\ & \quad f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2. \end{aligned}$$

It then suffices to show that $\Pi(\beta)$ is uniquely maximized at $\beta = \beta_0$. To this

end, write

$$\begin{aligned}
\Pi(\beta) &= \frac{1}{2} \int_{s_1 < s_2} g_\beta^*(|s_1 - s_2|) f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2 \\
&\quad - \frac{1}{2} \int_{s_1 > s_2} g_\beta^*(|s_1 - s_2|) f_{\tilde{Z}^*}(s_1) f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\tilde{Z}^*}(s_2) f_{\epsilon^*}(\tilde{t}_2 - s_2) ds_1 ds_2 \\
&= \frac{1}{2} \int_{s_1 < s_2} g_\beta^*(|s_1 - s_2|) f_{\tilde{Z}}(s_1) f_{\tilde{Z}}(s_2) \\
&\quad [f_{\epsilon^*}(\tilde{t}_1 - s_1) f_{\epsilon^*}(\tilde{t}_2 - s_2) - f_{\epsilon^*}(\tilde{t}_1 - s_2) f_{\epsilon^*}(\tilde{t}_2 - s_1)] ds_1 ds_2,
\end{aligned} \tag{7.3}$$

where we define $g_\beta^*(t) = P(|\xi(\beta)| < t)$ and then $g_{\beta_0}^* = 1$ since $\xi(\beta_0) \equiv 0$.

Since $g^*(\cdot)$ is only maximized at $\beta = \beta_0$ by assumption, to show that β_0 is the unique maximizer of $g(\beta)$, we only need to prove that the quantity in the square brackets is positive for all $\tilde{t}_1 < \tilde{t}_2$ and $s_1 < s_2$.

Now we show

$$h(\tilde{t}_1 - s_1) + h(\tilde{t}_2 - s_2) > h(\tilde{t}_1 - s_2) + h(\tilde{t}_2 - s_1)$$

for all $\tilde{t}_1 < \tilde{t}_2$ and $s_1 < s_2$, where $h = \log f_\epsilon$.

By the fact that f_{ϵ^*} is log-concave,

$$\frac{\partial}{\partial t} (h(t - s_1) - h(t - s_2)) = \int_{t-s_2}^{t-s_1} \frac{d^2}{ds^2} h(s) ds < 0.$$

Therefore $h(t - s_1) - h(t - s_2)$ is decreasing in t . As a result,

$$h(\tilde{t}_1 - s_1) + h(\tilde{t}_2 - s_2) > h(\tilde{t}_1 - s_2) + h(\tilde{t}_2 - s_1).$$

Step 2. We show that

$$\sup_{\beta} |g_n(\beta) - g(\beta)| = O_p\left(\sqrt{\frac{\log n}{n}}\right). \tag{7.4}$$

For each $n \in \mathcal{N}$, let $\{\beta_{n_1}, \dots, \beta_{n_m}\}$ be a $1/n^2$ -net of \mathbf{B} , which means that

$$\mathbf{B} \subset \cup_{k=1}^m B(\beta_{n_k}, \frac{1}{n^2}).$$

Then $m = O(n^{2d})$.

For $M > 1$, we have

$$\begin{aligned}
& P(\sup_{\beta} [g_n(\beta) - g(\beta)] > M\sqrt{\frac{\log n}{n}}) \\
& \leq P(\sup_{k=1, \dots, m} [g_n(\beta_{n_k}) - g(\beta_{n_k})] > (M-1)\sqrt{\frac{\log n}{n}}) \\
& \quad + P(\sup_{\beta} [g_n(\beta) - g(\beta)] - \sup_{k=1, \dots, m} [g_n(\beta_{n_k}) - g(\beta_{n_k})] > \sqrt{\frac{\log n}{n}}). \quad (7.5)
\end{aligned}$$

By Hoeffding's inequality (1963) for U-statistics, the first term in the right hand side of (7.5) can be bounded by $O(n^{2d-(M-1)^2/4})$. Using Chebyshev's inequality, the second term in the right hand side of (7.5) is bounded by $O(\frac{1}{n^2})$.

Now we have shown that

$$\begin{aligned}
& P(\sup_{\beta} [g_n(\beta) - g(\beta)] > M\sqrt{\frac{\log n}{n}}) \\
& = O(n^{2d-(M-1)^2/4}) + O(\frac{1}{n \log n}). \quad (7.6)
\end{aligned}$$

Since the last equality still holds if we replace g_n and g by $-g_n$ and $-g$, it can be written as

$$\begin{aligned}
& P(\sup_{\beta} |g_n(\beta) - g(\beta)| > M\sqrt{\frac{\log n}{n}}) \\
& = O(n^{2d-(M-1)^2/4}) + O(\frac{1}{n \log n}). \quad (7.7)
\end{aligned}$$

Then it follows equality (7.4).

Step 3. We show that $\hat{\beta}_n$ converges to β_0 in probability.

Since β_0 is the unique maximizer of g , and $\hat{\beta}_n$ is the maximizer of g_n , we have

$$\begin{aligned}
0 & \leq g(\beta_0) - g(\hat{\beta}_n) \\
& = [g(\beta_0) - g_n(\beta_0)] - [g(\hat{\beta}_n) - g_n(\hat{\beta}_n)] - [g_n(\hat{\beta}_n) - g_n(\beta_0)] \\
& \leq [g(\beta_0) - g_n(\beta_0)] - [g(\hat{\beta}_n) - g_n(\hat{\beta}_n)] \\
& = O_p(\sqrt{\frac{\log n}{n}}) + O_p(\sqrt{\frac{\log n}{n}}) \\
& = O_p(\sqrt{\frac{\log n}{n}}) \quad (7.8)
\end{aligned}$$

On the other hand, by the differentiability of density functions of \tilde{Z} and X , note that β_0 is the unique maximizer of g and $\dot{g}(\beta_0) = 0$, the Taylor expansion can then be written as

$$g(\hat{\beta}_n) - g(\beta_0) = -(\hat{\beta}_n - \beta_0)' A(\hat{\beta}_n - \beta_0) + o_p(\hat{\beta}_n - \beta_0)^2, \quad (7.9)$$

where A is a positive definite matrix.

Compare the last two equations, it follows that

$$\hat{\beta}_n - \beta_0 = O_p\left(\sqrt[4]{\frac{\log n}{n}}\right) = o_p(n^{-1/5}). \quad (7.10)$$

The consistency is proved.

Asymptotic normality:

We still use the notation of g and g_n as above. Furthermore, denote

$$\epsilon_n(\beta) = g_n(\beta) - g(\beta). \quad (7.11)$$

Standard decomposition of U-statistics gives

$$\epsilon_n(\beta) - \epsilon_n(\beta_0) = \frac{1}{n} \sum_{i=1}^n b_i(\beta) + \frac{1}{n^2 - n} \sum_{i < j} d_{ij}(\beta), \quad (7.12)$$

where

$$b_i(\beta) = E[a_{ij}(\beta) + a_{ji}(\beta) - 2Ea_{ij}(\beta)|Z_i, X_i, Y_i], \quad (7.13)$$

$$d_{ij}(\beta) = a_{ij}(\beta) + a_{ji}(\beta) - 2Ea_{ij}(\beta) - b_i(\beta) - b_j(\beta). \quad (7.14)$$

and

$$a_{ij}(\beta) = [I\{Z_i + \beta' X_i > Z_j + \beta' X_j\} - I\{Z_i + \beta_0' X_i > Z_j + \beta_0' X_j\}] I\{Y_i > Y_j\}. \quad (7.15)$$

Note that $Eb_i(\beta) \equiv 0$, Taylor expansion gives

$$\frac{1}{n} \sum_{i=1}^n b_i(\beta) = (\beta - \beta_0)' \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) + o_p(|\beta - \beta_0|)^2. \quad (7.16)$$

Using exponential inequality again, similar to the step 2 in the proof of consistency, we have

$$\sup_{|\beta - \beta_0| = o_p(n^{-1/5})} \left| \frac{1}{n^2 - n} \sum_{i < j} d_{ij}(\beta) \right| = o_p(n^{-1}). \quad (7.17)$$

So far we have shown that

$$\begin{aligned} & g_n(\beta) \\ = & g(\beta) + \epsilon_n(\beta) \\ = & g(\beta_0) - \frac{1}{2}(\beta - \beta_0)' A(\beta - \beta_0) + (\beta - \beta_0)' \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) + \epsilon_n(\beta_0) + o_p(|\beta - \beta_0|)^2 \\ & + o_p(n^{-1}) \\ = & f_n(\beta) + \epsilon_n(\beta_0) + o_p(n^{-1}), \end{aligned} \quad (7.18)$$

where

$$\begin{aligned} & f_n(\beta) \\ = & g(\beta_0) - \frac{1}{2}(\beta - \beta_0)' A(\beta - \beta_0) + (\beta - \beta_0)' \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) + o_p(|\beta - \beta_0|)^2 \\ = & g(\beta_0) - \frac{1}{2}(\beta - \beta_0)' A_n(\beta - \beta_0) + (\beta - \beta_0)' \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) \\ = & g(\beta_0) - \frac{1}{2} \{ A_n^{1/2} [\beta - \beta_0 - A_n^{-1} \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0)] \}' \{ A_n^{1/2} [\beta - \beta_0 - A_n^{-1} \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0)] \} \\ & + \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) \right)' A_n^{-1} \left(\frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) \right), \end{aligned} \quad (7.19)$$

where we let $o_p(|\beta - \beta_0|)^2 = c_n |\beta - \beta_0|^2$ with $c_n = o_p(1)$ and $A_n = A - 2c_n I$.

So the maximizer of f_n is

$$\hat{\gamma}_n = \beta_0 + A_n^{-1} \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) \quad (7.20)$$

Suppose that $\hat{\beta}_n$ is the maximizer of g_n , then

$$\begin{aligned}
0 &\leq f_n(\hat{\gamma}_n) - f_n(\hat{\beta}_n) \\
&= [f_n(\hat{\gamma}_n) + \epsilon_n(\beta_0) - g_n(\hat{\gamma}_n)] - [f_n(\hat{\beta}_n) + \epsilon_n(\beta_0) - g_n(\hat{\beta}_n)] - [g_n(\hat{\beta}_n) - g_n(\hat{\gamma}_n)] \\
&\leq [f_n(\hat{\gamma}_n) + \epsilon_n(\beta_0) - g_n(\hat{\gamma}_n)] - [f_n(\hat{\beta}_n) + \epsilon_n(\beta_0) - g_n(\hat{\beta}_n)] \\
&= o_p(n^{-1}) + o_p(n^{-1}) \\
&= o_p(n^{-1}).
\end{aligned} \tag{7.21}$$

On the other hand, from the expression of f_n ,

$$\begin{aligned}
&f_n(\hat{\gamma}_n) - f_n(\hat{\beta}_n) \\
&= \frac{1}{2} \{A_n^{1/2}[\hat{\beta}_n - \beta_0 - A_n^{-1} \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0)]\}' \{A_n^{1/2}[\hat{\beta}_n - \beta_0 - A_n^{-1} \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0)]\}.
\end{aligned} \tag{7.22}$$

Compare (7.21) and (7.22), finally we have

$$\begin{aligned}
\hat{\beta}_n &= \beta_0 + A_n^{-1} \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) + o_p(n^{-1/2}) \\
&= \beta_0 + A^{-1} \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) + (A_n^{-1} - A^{-1}) \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) + o_p(n^{-1/2}) \\
&= \beta_0 + A^{-1} \frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) + o_p(n^{-1/2}),
\end{aligned} \tag{7.23}$$

where the last equation comes from that

$$A_n^{-1} - A^{-1} = o_p(1)$$

and

$$\frac{1}{n} \sum_{i=1}^n \dot{b}_i(\beta_0) = O_p(n^{-1/2})$$

by the definition of A_n and the central limit theorem.

Therefore,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = A^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{b}_i(\beta_0) + o_p(1) \rightarrow N(0, \Sigma)$$

in distribution, where

$$\Sigma = A^{-1}Var\{\dot{b}_1(\beta_0)\}(A^{-1})'.$$

Table 7.1: *Simulation results for different models and sampling schemes.*

S	Proposed							IP						
	BIAS	SE	SEE	CP	BIAS	SE		BIAS	SE	SEE	CP	BIAS	SE	
1	$\epsilon \sim \text{Double Exponential}$							$\epsilon \sim \text{Extreme Value}$						
	β_1	0.0238	0.2213	0.2286	0.97	0.3004	0.0966	0.0390	0.1443	0.1632	0.98	-0.2433	0.1031	
	β_2	0.0030	0.2337	0.2320	0.99	-0.3617	0.1459	0.0234	0.1339	0.1638	0.97	-0.2039	0.1438	
	$\epsilon \sim \text{Normal}$							$\epsilon \sim \text{Logistic}$						
	β_1	0.0159	0.1338	0.1368	0.98	0.2149	0.0540	0.0084	0.1933	0.1826	0.94	0.5993	0.0918	
	β_2	0.0234	0.1430	0.1356	0.97	-0.2661	0.0603	0.0041	0.2117	0.1901	0.92	-0.3347	0.1376	
2	$\epsilon \sim \text{Double Exponential}$							$\epsilon \sim \text{Extreme Value}$						
	β_1	0.0142	0.1887	0.1816	0.96	0.3069	0.0692	0.0169	0.0933	0.0968	0.96	0.0732	0.0389	
	β_2	0.0216	0.1783	0.1821	0.98	-0.1738	0.0887	0.0149	0.0963	0.0967	0.96	-0.0304	0.0461	
	$\epsilon \sim \text{Normal}$							$\epsilon \sim \text{Logistic}$						
	β_1	0.0020	0.1469	0.1443	0.95	0.2015	0.0490	0.0155	0.2835	0.2755	0.98	0.4789	0.0739	
	β_2	0.0006	0.1442	0.1431	0.97	-0.1217	0.0600	-0.0256	0.3052	0.2656	0.98	-0.2841	0.0993	
3	$\epsilon \sim \text{Double Exponential}$							$\epsilon \sim \text{Extreme Value}$						
	β_1	0.0161	0.1426	0.1394	0.96	0.1556	0.1198	0.0087	0.0649	0.0801	0.99	-0.0620	0.2188	
	β_2	0.0155	0.1593	0.1384	0.96	-0.0297	0.1683	0.0032	0.0786	0.0811	0.98	0.1327	0.3064	
	$\epsilon \sim \text{Normal}$							$\epsilon \sim \text{Logistic}$						
	β_1	0.0067	0.0989	0.0957	0.94	0.0364	0.1554	0.0354	0.2108	0.1931	0.99	0.1795	0.1376	
	β_2	0.0017	0.0880	0.0937	0.99	-0.0249	0.1484	0.0239	0.1986	0.1870	0.98	-0.0167	0.1744	
4	$\epsilon \sim \text{Double Exponential}$							$\epsilon \sim \text{Extreme Value}$						
	β_1	0.0306	0.0784	0.0874	0.96	0.0035	0.0746	0.0272	0.0825	0.0858	0.96	-0.2608	0.0681	
	β_2	0.0036	0.0708	0.0887	0.99	-0.0072	0.0851	0.0031	0.0717	0.0864	0.99	0.1630	0.0786	
	$\epsilon \sim \text{Normal}$							$\epsilon \sim \text{Logistic}$						
	β_1	0.0028	0.0948	0.0797	0.96	0.0035	0.0527	0.0245	0.1282	0.1263	0.98	0.1689	0.0844	
	β_2	0.0014	0.0794	0.0794	0.96	-0.0054	0.0615	0.0105	0.1005	0.1225	0.99	-0.0147	0.1011	
5	$\epsilon \sim \text{Double Exponential}$							$\epsilon \sim \text{Extreme Value}$						
	β_1	0.0070	0.2447	0.2464	0.97	0.0995	0.0760	0.0451	0.2478	0.2332	0.97	-0.1797	0.0712	
	β_2	0.0206	0.2971	0.2496	0.96	-0.1329	0.0843	0.0081	0.2227	0.2381	0.97	-0.0206	0.0880	
	$\epsilon \sim \text{Normal}$							$\epsilon \sim \text{Logistic}$						
	β_1	0.0262	0.4074	0.3964	0.96	0.1311	0.0496	0.0417	0.2045	0.2083	0.98	-0.0349	0.0832	
	β_2	0.0402	0.4028	0.3907	0.96	-0.1645	0.0583	0.0389	0.1942	0.2082	0.95	0.2748	0.0952	

Note: S represents sampling scheme; IP represents the inverse probability method.